Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 227-230 recibido 08-04-2013 revisado 16-06-2013 aceptado 21-06-2013

TIMPANO: Technology for complex Human-Machine conversational interaction with dynamic learning

TIMPANO: Tecnología para interacción conversacional hombre-máquina compleja con aprendizaje dinámico.

Emilio Sanchis
ELiRF-Universitat Politecnica Valencia
Camino de Vera s/n 46022 Valencia
esanchis@dsic.upv.es

Alfonso Ortega
Universidad de Zaragoza
C/ María de Luna 1. 50018 Zaragoza
ortega@unizar.es

M. Inés Torres
Universidad del País Vasco UPV/EHU
Campus de Leioa. 48940 Leioa. Bizcaia
manes@we.lc.ehu.es

Javier Ferreiros
Universidad Politécnica Madrid
Ciudad Universitaria s/n. Madrid
jfl@die.upm.es

Resumen: El proyecto TIMPANO tiene por objetivo profundizar en el desarrollo de sistemas de comunicación oral hombre-máquina atendiendo principalmente a la capacidad de dar respuesta a múltiples requerimientos de los usuarios, como pueden ser el acceso a información, la extracción de información, o el análisis de grandes repositorios de información en audio. En el proyecto se hace especial énfasis en la adaptación dinámica de los modelos a diversos contextos, tanto de tipo acústico, como semántico o de idioma.
Palabras clave: Tecnologías del Habla, Interacción oral.

Abstract: The goal of the TIMPANO project is to research about the development of speech-driven human-machine interaction systems, regarding mainly the ability of answering multiple requirements from the users, such as accessing and extracting information, and analyzing large repositories of audio information. This project is especially focused in the dynamic adaptation of the models to different acoustic and semantic contexts as well as to different languages.
Keywords: Speech Techonogies, Oral interaction.

1 Informations of the Project

This Project is founded by the "Ministerio de Economía y Competitividad" TIN2011-28169-C05 and there are four reseach groups involved in it: ELiRF (Universidad Politécnica de Valencia), ViVoLab (Universidad de Zaragoza), TR&ST (Universidad del Pais Vasco), GTH (Universidad Politécnica de Madrid).

2 Introduction

Significant advances have been achieved in speech technologies allowing researchers to face new challenges in different areas of human-machine interaction slightly exploited so far, improving and consolidating already developed technologies at the same time. The new generation of speech interaction systems requires a progressive increase of robustness against the variability of users, languages and acoustic conditions, as well as against the huge volume of information those systems provide access to. The international scientific community is devoting strong efforts to provide solutions allowing to face unconstrained vocabularies, different tasks, several languages and open domain information retrieval systems for which an exhaustive signal analysis helps to extract as much information as possible.

Speech technologies, like some other disciplines that have arisen based on the demands of the information society, should

ISSN 1135-5948

have a major role in the future of the management and use of the large amount of digital contents that are being generated. The increasing flow of information in different languages and through different media, the large audio repositories (conferences, parliaments, broadcasting, ..), interactive information services, etc. are some examples of the new challenges the scientific community is facing. But this requires broadening the concept of oral interaction; because it is not just a problem of obtaining or emitting a sequence of words, but also detecting voice activity, who spoke in a given moment, in which language, the semantic information contained, etc. to be able to seek certain segments, or cooperate with the machine to achieve the proposed goal. Human speech is not only characterized by segmental information and spectral correlates of each sound. Several cognitive, emotional and contextual factors affect speech variability which operates at higher levels than phonetics. This higher level variability affects several speech features, including sound intensity, intonation or stops timing, and bears consequences on the emotional state of speech. Human beings do make a successful intensive use of this prosodic information to favor our communication processes. Thus, incorporating prosodic processing capabilities to advanced spoken interaction systems is a key factor to improve interaction results.

As a consequence of facing this problems, where the volume of information as well as the variety of users, languages, or media is enormous, the relevance of machine learning methods to allow systems to dynamically adapt to the new features of the interaction, or to detect and model new events (new words, concepts, users,..) is really important. Since obtaining these kind of systems from a-priori information or with models learned from few samples is unfeasible, it is necessary for the system to have self-learning mechanisms based on the evaluation of its behavior or, at least, ways to generate information to be readily used by the designer for adapting the models dynamically, and with little effort (in the line to provide tools to help designers).

In this framework, the research project proposed here is oriented to the development of voice interaction systems able to address new and more complex tasks, both in the sense of working with larger amounts of data and increasing the complexity of the application domain. Obviously we bound the specific areas where we are going to focus our work. Moreover, we try to follow a realistic approach, encouraging the development of assistive tools that for certain applications can facilitate the fulfillment of specific tasks by using spoken interaction systems. Much of the efforts will be focused on developing and applying self-learning techniques at different levels of knowledge representation, maximizing the benefit of using the data generated through the interaction with the system. On the other hand, special attention will be paid to the development and improvement of methodologies to process large amounts of audio information.

As a summary, it is worth to highlight the following lines of work in this project:

- Development of conversational interactive systems along two different lines: increase the robustness of dialogue systems for restricted domains, and facilitate the access to large amounts of audio data, improving two-way human-machine communication process (Griol et al., 2008) (Ortega et al., 2010).
- Performance improvement of several technological components needed for human-machine interaction.
    - Processing, analysis and distillation of information from raw data: speaker and language identification, acoustic segmentation, event detection, speech recognition for open-vocabulary tasks, topic detection, content retrieval… (Guijarrubia y Torres, 2010) (Pardo, Ferreiros y Montero, 2011)
    - Propose new learning techniques to allow dynamic model adaptation at all levels of knowledge (acoustic, lexical, semantic and dialogue) (Buera et al., 2010).
    - Deal with other components of conversational interaction as prosody, emotion and multilingualism (Perez-Ramirez, Torres y Casacuberta, 2008) (Barra-Chicote et al., 2010).

- In certain tasks give higher priority to systems that assist the user versus the fully automatic ones. Tasks that would be impossible to deal with due to the large amount of information to be processed can become tractable thanks to the use of systems that do part of the work.
- Regarding the development of applications we propose two different goals.
    - On one hand develop tools where conversational interactive systems assist the access to both structured and unstructured information,
    - On the other hand to take advantage of speech technologies on the design of socially useful applications such as systems to help people with special needs (i.e., speech/language disorders, physical disabilities like visual or hearing impairments …) Examples of this kind of systems are tools for speech therapy, aid devices for deaf people or sign language translation, areas in which some groups of the consortium have proven experience (Saz et al., 2010).

3 Objetives of the Project

Strategic Objectives:

This project will focus on covering gaps identified in existing systems of spoken human-machine interaction in areas such as:

- **Self-assessment** systems that provides information about the quality of their results.
- **Self-learning**, which allows systems to improve their use and adapt easily to new situations.
- **Scalability** and extensive interaction domains, connected with the need to address open-vocabulary tasks and large and diverse information repositories.

- Its usefulness as assistive tools to allow the user the fulfillment of tasks whenever they become complex or in situations when the user has serious difficulties (certain disabilities or speech and language disorders).

Scientific-Technological objectives:

These objectives are divided into three fundamental aspects: the treatment of acoustic information, treatment of the interaction and the transition from the acoustic domain to the interaction domain and vice versa.

- In terms of the acoustic aspects, the researchers aim to develop modules such as the identification / classification of acoustic events, segmentation / diarization / clustering or automatic speaker or language recognition all of them being able to offer, along with its outputs, an indication of how reliable these are, ie the capacity of self-assessment, using the information obtained from all the available sources. Also, if the self-assessment is not positive, we propose to exploit information obtained through user interaction to allow the modules to evolve as independently as possible.

- With regard to the processing of the interaction, to advance the semantic modeling by using unsupervised learning with automatic detection of new semantic events. Progress in the dynamic learning interaction systems in restricted domains, autonomously adapting them to specific tasks and modeling the temporal evolution of personalized interaction. In open domains, to assist the user in accessing unstructured information. Wherever necessary the treatment of multilingualism will be addressed, and if necessary, translation tools to assist the interaction will be used. Progressing this kind of systems to incorporate the emotional component in the interaction with the detection of user's emotional state, combination with an internal emotional model and appropriate response generation (emotional speech synthesis).

- In the transition between both domains, we will mainly develop two aspects, acoustic modeling and language modeling. According to the first one, we propose improvements in statistical modeling of acoustic emissions to be included in automatic speech recognition systems (ASR), modeling of disorders and variations of speech, the use of language independent acoustic units or robust feature extraction and specific classifiers: tandem, or hybrid. The on-line learning to adapt to speakers (including to those unregistered), speech pathologies, language / dialect / accent and also advances in acoustic modeling, also

for personalized speech synthesis. Regarding language modeling, progress in large vocabulary ASR, in the detection of terms and words out of vocabulary and progress in the adaptability to unsupervised semantic domain of finite-state transducers.

Implementation Objectives

In this field, this project proposes the development of tools that enhance the technological capabilities of the consortium, progressing the software architecture which is already available from previous common projects and the building of prototypes and demonstrators which show the achievements in this project:

- Adaptable Interaction Systems in Restricted Domains.
- Assistive Interaction Systems Including People with Special Needs (e-inclusion).
- Support to the Processing of Multimedia and Unstructured Information.

4 Aknowledgements

This work is founded by the "Ministerio de Economía y Competitividad" TIN2011-28169-C05

References

Barra-Chicote Roberto, Yamagishi Junichi, King Simon, Montero Juan Manuel, Macías-Guarasa Javier . 2010 "Analysis of Statistical Parametric and Unit-Selection Speech Synthesis Systems Applied to Emotional Speech" *Speech Communication*, Volume 52 Issue 5, Pages 394-404.

Buera, Luis, A. Miguel, O. Saz, A. Ortega y E. Lleida Solano. 2010 "Unsupervised Data-Driven Feature Vector Normalization With Acoustic Model Adaptation for Robust Speech Recognition*" IEEE Transactions on Audio, Speech and Language Processing" . Vol.* 18, no. 2, pp. 296-309.

Griol David, Hurtado Lluís F.,Segarra Encarna, y Sanchis Emilio. 2008. Acquisition and Evaluation of a Dialog Corpus throughWOz and Dialog Simulation Techniques. *Speech Communication*, 50, pp. 666-682.

Guijarrubia, Víctor G. y Torres, M. Inés. 2010, "Text- and speech-based phonotactic models for spoken language identification of Basque and Spanish", *Pattern Recognition Letters*, 31, 6: 523 - 532,

Ortega Lucia, Galiano Isabel, Hurtado Lluís-F., Sanchis Emilio y Segarra Encarna. 2010. A Statistical Segment-Based Approach for Spoken Language Understanding. Proc. of INTERSPEECH'10, pp. 1836-1839.

Pardo, J.M., Ferreiros, J. y Montero, J.M. 2011 "Speaker diarization based on intensity channel contribution", *IEEE Transactions on Audio, Speech and Language Processing* 19 , Vol.19, Issue: 4 Pages: 754 – 761.

Pérez-Ramírez, A., Torres, M. Inés y Casacuberta, F. 2008, "Joining linguistic and statistical methods for Spanish-to-Basque speech translation", *Speech Communication*, 50: 1021-1033.

Saz O., Yin S.-C., Lleida E., Rose R., Vaquero C. y Rodriguez W. R.. 2009. "Tools and technologies for Computer-Aided Speech and Language Therapy". *Speech Communication*, Special Issue on Spoken Language Technology for Education.